Proceedings of CONF-SEML 2026 Symposium: Importance of Machine Learning Methods and Analysis in Engineering
DOI: 10.54254/2755-2721/2026.AS32265

Predicting Exam Success Using Machine Learning: An Analysis of Learning Behaviors

Yaoyang Huang

Xi'an Jiaotong-Liverpool University, Suzhou, China
Yaoyang.Huang@student.xjtlu.edu.cn

Abstract. The availability of learning activity data from online education platforms has created new opportunities to examine how student behaviors relate to academic outcomes. Within this context, educational data mining has been widely applied to analyze learning patterns and support performance prediction. This paper explores whether students' learning behaviors can be used to predict exam success in a Python learning environment. Exploratory data analysis is used to compare behavioral characteristics between students who passed and those who failed the exam. The prediction task is formulated as a binary classification problem using the variable passed exam. A support vector machine (SVM) classifier is applied to distinguish between pass and fail outcomes, and feature importance analysis is conducted to better understand the contribution of different learning behaviors. The results suggest that engagement-related variables, particularly study time and practice activities, are closely associated with exam success, while demographic features contribute relatively little to prediction performance. These findings are consistent with existing educational data mining research and demonstrate the value of machine learning methods for analyzing learning behavior data.

Keywords: Machine Learning, Educational Data Mining, Classification, Learning Behavior, Exam Performance

1. Introduction

The rapid expansion of online learning has led to the generation of large volumes of educational data that record students' learning activities, engagement levels, and assessment results. As a result, educational data mining has attracted increasing attention as a means of extracting useful insights from such data and improving learning and teaching practices [1,2]. Previous research has shown that machine learning techniques can be effective tools for predicting student academic performance and identifying factors associated with learning success [3,4].

One important application of educational data mining is the prediction of exam outcomes. Being able to identify students who are at risk of failing at an early stage allows educators to provide timely academic support and adjust instructional strategies [5,6]. Earlier studies indicate that variables describing learning behaviors, such as time spent studying or participation in practice activities, often provide stronger predictive power than demographic characteristics [4,7]. For instance, Cortez and Silva [3] demonstrated that data mining approaches can successfully model

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.AS32265

student performance, while Kotsiantis et al. [4] reported strong results using classification techniques in distance learning environments.

Support vector machine (SVM) has been widely used in classification problems due to its robustness and ability to handle complex decision boundaries. In educational contexts, Huang and Fang [8] showed that SVM performs well when predicting student academic performance, making it a suitable choice for this type of task. Building on this line of work, the present study applies an SVM-based approach to learning behavior data collected from a Python learning context.

The aim of this study is to determine if the variables of learning behavior performance affect whether a student can pass or fail an examination. Through exploratory data analysis and machine learning classification, this study aims to determine which behaviors are closely related to passing an examination.

2. Dataset and problem defining

The dataset of this work contains detailed student learning behaviors and exam score records in the context of learning Python. A variable used for each student is the weekly learning engagement. Some of the components of study engagement are study time per week, number of practice problems solved, completed learning projects, and self-reported confidence. These variables are common indicators used in educational data mining research given that they showcase student effort, persistence and interaction with learning materials [1,2].

Final-exam score and passed-exam measures are behaviors related to the exam in addition to studying behaviors. The final-exam score variable measures students' numeric score in the last exam. The binary variable passed_exam, which indicates whether the exam was passed or failed, is derived from the score using a threshold. In many educational contexts, pass-fail outcomes are the most important criteria for decisions about academic progression and intervention. Thus, passed_exam can constitute a realistic and practically meaningful indicator of academic success [3,6].

Choosing passed_exam as the target variable, the problem can be formulated as a binary classification problem from a methodological perspective. As regression models can be used to predict continuous exam scores, classification models are commonly employed when the goal of the research is to identify students at risk of failing or support an early-warning system in education [5,9]. It simplifies interpretation and matches factor instructors and institutions use to make decisions.

Educational data is often noisy, accompanied by behavioral variance and nonlinear links between actions and grades. Because of such characteristics, an incorporation of classification-based approaches, especially those using different machine learning algorithms capable of modeling complex decision boundaries works well. This means that modeling the problem as a pass/fail classification task will help us to arrange a clear way for further analysis and modeling [4].

3. Exploring behaviors of the learning process

An exploratory data analysis (EDA) was performed to obtain an overall understanding of the data and also how the learning behaviors of those students who passed the exam and failed the exam are different. In educational data mining, exploratory data analysis (EDA) is very useful as it reveals underlying patterns, and examines potential problems such as class imbalance, as well as helps to choose features for modeling [2].

DOI: 10.54254/2755-2721/2026.AS32265

The first observation of the target variable distribution indicates that there is a clear imbalance between classes pass and fail, evident from the distribution of failures among students (Figure 1). It is a common characteristic of educational datasets to have classes that are sometimes imbalanced [7]. At an early stage, we recognize the data imbalance issue. Thus, it demands a careful interpretation of results, and possibly the utilization of evaluation metrics aside from mere accuracy.

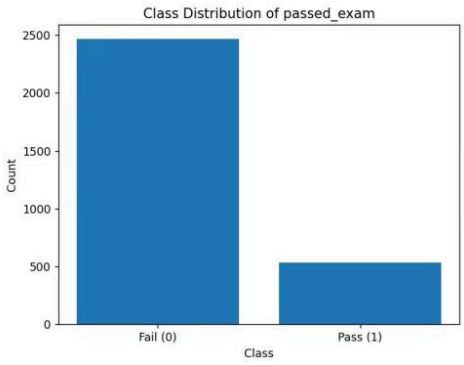


Figure 1. Distribution of pass and fail outcomes in the dataset

There are considerable differences between the two outcome groups' weekly study time analysis (Figure 2). Typically, students who passed the exam report more study hours per week, while those who failed are more concentrated at lower values of study time. While their distributions have a lot of overlap, the fact that there is an overall shift upwards in the passing students means spending more time on the exam question paper is associatively related to passing. The finding conforms to the studies that stress the continuous effort put into study for achievement [4,10].

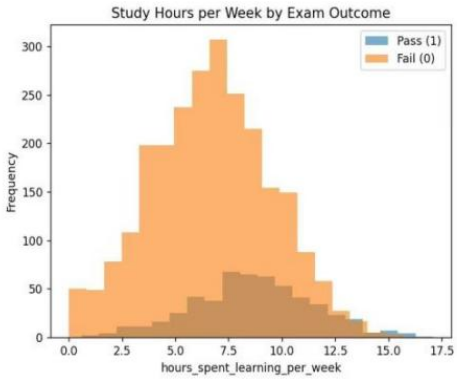


Figure 2. Distribution of weekly study time for students who passed and failed the exam

There is a similar pattern when assessing the number of practice problems. Students tend to pass the examination, and complete more practice activities. More practice activities on the learning materials indicate more activity. Baker and Inventado [1] and Peña-Ayala [7] observe that there is a major recognition of practice-based learning in acquiring skills particularly, in programming-related subjects where experience of solving problems matters. The differences found between pass and fail groups may hint that practice intensity is an important predictor of exam results (Figure 3).

DOI: 10.54254/2755-2721/2026.AS32265

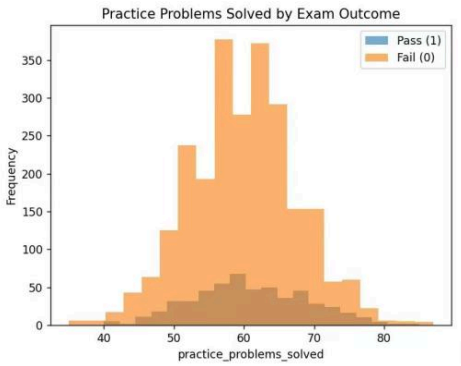


Figure 3. Distribution of practice problems solved for pass and fail group

A correlation analysis was subsequently carried out on all numerical attributes (Figure 4). Based on the analysis, engagement-related mechanisms, such as study time and practice activities, are more strongly associated with exam outcomes than demographic and self-reported mechanisms. The data mining in education has shown that student performance is largely predictable from behavioral indicators rather than from background characteristics. Romero and Ventura [2] established that many previous studies of educational data mining confirm this fact. Justification for focusing on the learning behavior features in the classification model has been provided.

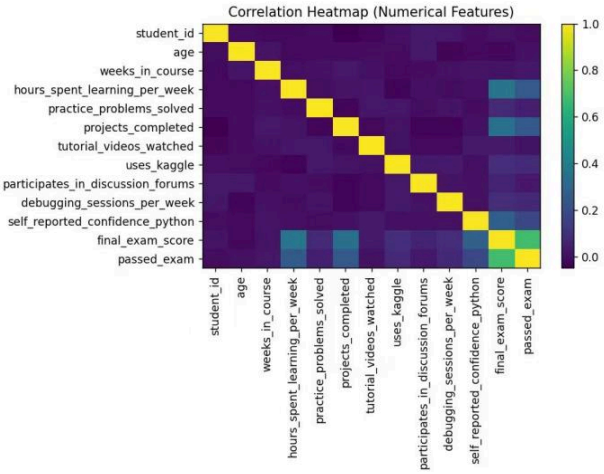


Figure 4. Correlation heatmap of numerical learning behavior features and exam outcomes

4. Machine learning methodology for prediction of exam results

After conducting an exploratory data analysis, the research team selected the learning behavior variables as input features for the model. Educational data mining feature selection refers to the efforts that strike a balance between predictive performance and interpretability so that the imperative model outputs can be related to learning behavior and educational practice meaningfully [7].

Before the training of the model, the preprocessing techniques were applied on the data. The procedures had to deal with variable formats and ensure that features were appropriate for an ML algorithm. A stable model performance can only be achieved by decoupling irrelevant variation in the data through such preprocessing [4].

DOI: 10.54254/2755-2721/2026.AS32265

An SVM classifier was built for predicting pass or fail for a student for the exam. The study by Huang & Fang [8] states that SVM is a most-used machine learning algorithm for binary classification. Further, it has performed well on education performance prediction. One benefit of SVM is that it is able to model complex, nonlinear relationships between input features and outcomes using kernel functions. This can be particularly suitable for educational settings, where learning behaviors may affect each other in non-obvious ways.

Also, SVM is known to be effective even when the number of features is much greater than the number of samples. In other words, SVM is robust in high-dimensional feature space and resists overfitting. Due to these abilities, it is essential to use such a tool on datasets with multiple behavioral indicators, as is seen in learning analytics [9]. To capture useful learning patterns from the student population for pass and fail cases, the SVM classifier aims to determine a suitable decision boundary between them.

5. Analysis of feature importance

For enhanced interpretability, a feature importance analysis was performed to analyze the contributions of the different learning behaviors to the classification. SVM models are typically considered to be less interpretable than simpler models. However, feature importance techniques can sometimes provide useful insights as to which features are most influential for any modelling [8].

As the research shows, high study time and practice-related variables are the most distinguished factors that make a student pass or fail (Figure 4). As a result, they support our exploratory analysis and suggest that ongoing and active engagement may be a significant factor in academic success. The educational data mining studies show similar outcomes, as the predictive value of effort-based indicators was stressed in these studies [4,10].

Comprehending the importance of features has practical applications in education. Through identifying the learning behavior patterns that are most highly associated with exam performance, interventions can be planned and effective studying encouraged. Through the lens of learning analytics, feature importance analysis can link predictive modeling with actionable insights for use in education [6].

6. Discussion

According to the findings of the study, they match the results based on educational data mining literature, where learning behavior has been shown to be an important predictor of performance [2,7]. The observed regular impact of study time and practice activity in exploratory analysis and model interpretation indicates that engagement-related variables play a decisive role in determining exam success.

In addition, the adoption of the binary classification framework captures the real-life scenario of educational decision-making, where defining pass or fail is often more relevant than the absolute score. Similar approaches have been applied successfully in early-warning systems for identifying academically at-risk students [6,9].

Recognition of several limitations despite their contributions. The existence of class imbalance can hinder classification performance and interpretation. Consequently, the binary version of outcomes results in information loss relative to that of predicting scores. The dataset has a specific learning context, which limits its generalizability to other subjects or contexts [5].

DOI: 10.54254/2755-2721/2026.AS32265

7. Conclusion

Within the educational data mining framework, this study used machine learning techniques to classify students' exam outcomes based on their learning behaviors. The exploratory data analysis clearly revealed differing behaviors among students who passed and failed the exam. Specifically, study time and consolidation activities were consistently associated with exam results. Most students who passed showed greater engagement as compared to those of the students who failed. Students who failed tend to participate in activities. These findings highlight how behavior predictors can be effective indicators of achievement performance. Focusing on a practical and intuitive measure of success, we formulate the task as a binary classification problem where the variable tells whether the student passed or not. The application of a support vector machine classifier indicated that learning behavior variables can differentiate pass and fail. Conclusion outcome of the study evidence from feature importance analysis suggested that engagement-related variables, in particular study time and practice intensity, played a key role in the classification results. These results confirm previous studies in educational data mining, which find that engagement and learning effort are predictive of student performance. The class imbalance of the dataset may affect the performance of classification models and their interpretation. In addition, the use of binary literals for exam outcomes condenses information contained in (continuous) exam scores. In addition, the findings may not be generalizable due to the specific learning context of the dataset. Future researchers may assess more classification algorithms, use techniques to mitigate class imbalance, and add alternate datasets. Classifications may be enhanced with such extensions for more reliable results and greater behavioral understanding and academic performance.

References

[1] Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In Learning Analytics. Springer, pp. 61-75.
[2] Bowers, A. J., & Sprott, R. (2012). Why tenth graders fail to complete high school. Review of Educational Research, 82(2), 231-269.
[3] Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In Proceedings of the 5th Future Business Technology Conference (FUBUTEC).
[4] Huang, S., & Fang, N. (2013). Predicting student academic performance using SVM. International Journal of Engineering Education, 29(5), 1188-1195.
[5] Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students. Journal of Learning Analytics, 1(1), 6-47.
[6] Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2012). Predicting students' performance in distance learning using machine learning techniques. Applied Artificial Intelligence, 18(5), 411-426.
[7] Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
[8] Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications, 41(4), 1432-1462.
[9] Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40(6), 601-618.
[10] Shaleena, K. P., & Paul, S. (2015). Data mining techniques for predicting student performance. International Journal of Computer Applications, 107(15), 23-28.